



# Analysis of chloroplast genome characteristics and codon usage bias in 14 species of Annonaceae

Xiang Hu<sup>1</sup> · Yaqi Li<sup>2</sup> · Fuxuan Meng<sup>1</sup> · Yuanjie Duan<sup>1</sup> · Manying Sun<sup>1</sup> · Shiyong Yang<sup>1</sup> · Haigang Liu<sup>1</sup>

Received: 6 March 2024 / Revised: 18 May 2024 / Accepted: 21 May 2024 / Published online: 27 May 2024  
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2024

## Abstract

For the study of species evolution, chloroplast gene expression, and transformation, the chloroplast genome is an invaluable resource. Codon usage bias (CUB) analysis is a tool that is utilized to improve gene expression and investigate evolutionary connections in genetic transformation. In this study, we analysed chloroplast genome differences, codon usage patterns and the sources of variation on CUB in 14 Annonaceae species using bioinformatics tools. The study showed that there was a significant variation in both gene sizes and numbers between the 14 species, but conservation was still maintained. It's worth noting that there were noticeable differences in the IR/SC sector boundary and the types of SSRs among the 14 species. The mono-nucleotide repeat type was the most common, with A/T repeats being more prevalent than G/C repeats. Among the different types of repeats, forward and palindromic repeats were the most abundant, followed by reverse repeats, and complement repeats were relatively rare. Codon composition analysis revealed that all 14 species had a frequency of GC lower than 50%. Additionally, it was observed that the proteins in-coding sequences of chloroplast genes tend to end with A/T at the third codon position. Among these species, 21 codons exhibited bias ( $RSCU > 1$ ), and there were 8 high-frequency (HF) codons and 5 optimal codons that were identical across the species. According to the ENC-plot and Neutrality plot analysis, natural selection had less impact on the CUB of *A. muricate* and *A. reticulata*. Based on the PR2-plot, it was evident that base G had a higher frequency than C, and T had a higher frequency A. The correspondence analysis (COA) revealed that codon usage patterns different in Annonaceae.

**Keywords** Annonaceae · Chloroplast genome · Codon usage bias · Optimal codon · Natural selection

## Introduction

Chloroplast is an organelle for photosynthesis and metabolic reactions in plants, it contains specific genomes (Chakraborty et al. 2020). The chloroplast genome is much smaller than the nuclear genome (Wilson et al. 2011), and has the characteristics of moderate molecular weight, easy

sequencing, multiple copies, simple structure (Zhang et al. 2023), moderate nucleotide replacement rate of DNA, significant differences in molecular evolution rate between coding and non-coding regions, and good collinearity among various groups of chloroplast genomes (Parmar et al. 2022). The chloroplast has a set of maternally inherited chloroplast genomes independent of the nuclear genome (Sadhu et al. 2023). It has been widely used in plant species identification, genetic diversity analysis, and phylogenetic analysis (Finkeldey and Gailing 2013). In plant phylogenetics studies with the integration of phylogenetic and genomics, the advantage of chloroplast genome had gradually emerged, an increasing number of plant chloroplast genomes had been sequenced and assembled (Wanichthanarak et al. 2023), a great deal of genetic resources had been analyzed by different bioinformatics methods, it provides a new thought to solution the groups of classifying difficult (Simmonds et al. 2021). 61 codons are matching 20 amino acids in nature, due to the degeneracy of the genetic code, most amino acids

Xiang Hu and Yaqi Li have contributed equally to this work and share first authorship.

✉ Haigang Liu  
nkyhlg@yaas.org.cn

<sup>1</sup> Tropical Eco-agriculture Research Institute, Yunnan Academy of Agricultural Sciences, Yuanmou, Yunnan 651300, China

<sup>2</sup> Tropical and Subtropical Cash Crops Research Institute, Yunnan Academy of Agricultural Sciences, Baoshan, Yunnan 678000, China

are encoded by multiple codons, except methionine (Met) and tryptophan (Trp) by a single codon, and these codons encode the same amino acid are called synonymous codons (Li et al. 2016). In the absence of genetic mutation and natural selection, the synonymous codons should be used at the same frequency. Nevertheless, the synonymous codons of a particular amino acid may not be used at the same frequency, the phenomenon of codons being used unevenly is called CUB (Dehlinger et al. 2021). Now, the recognized hypothesis is that CUB is the result of a combination of three evolutionary forces: mutation, selection, and genetic drift (Morton 2022). In the current study, CUB has been detected in all species, and it has fundamental differences among the species with distant phylogenetic relationships (Somaratne et al. 2019). Therefore, using CUB in different species and selecting appropriate codons for transgene studies can improve the expression efficiency of exogenous genes and lay a foundation for subsequent studies on the molecular mechanisms of plant adaptation to environmental changes (Almutairi 2021). Numerous studies have confirmed that the applicability of CUB to the chloroplast genome level in plants. Therefore, the study of protein coding sequence characteristics and codon usage patterns in chloroplast genome will provide basic information for elucidating the phylogenetic relationship between plant species and chloroplast gene engineering.

Annonaceae is a primitive taxon closely related to angiosperms and is a prevalent family in tropical flora, it is renowned for its significant economic and medicinal importance. The largest family of Magnoliales is Annonaceae, which has approximately 129 genera and over 2300 species (Jiang and Li 1979). Except for *Asimina Adans* was found in temperate regions, the rest genera were found in tropical and subtropical regions of the world and it was an important component of the ecosystem in lowland tropical rainforests. Annonaceae plants are also widely distributed in China, they are mainly distributed from Southwest to Taiwan, most of them are produced in South China, and a few are distributed in East China. Annonaceae has been identified as having 24 genera, 103 species, and 6 varieties of Annonaceae (Li and Michael 2012). Most of the fruit of this family is edible, and the main cultivated species are *Annona squamosa* L., *Annona muricata* L., and *Annona reticulata* L. (Nogueira et al. 2022). The wild species include the *Fissistigma uonicum* in Fujian Province, the *Fissistigma oldhamii* in Ganzhou region of Jiangxi Province, the *Uvaria microcarpa* and *Uvaria grandiflora* in Guangdong and Guangxi region, etc., which were mostly picked and eaten by local people or plant lovers (Lei et al. 2022). Although the resources of the Annonaceae are abundant on the whole, the studies on its adaptive evolution and phylogeny have been limited, so the species, population size and distribution, nutritional

composition, and market demand of wild fruits of the Annonaceae have not been clearly understood, and it has not been effectively exploited and utilized.

In recent years, correlative studies have defined the overall phylogenetic relationships of the Annonaceae, but many are based on multiple plasmid DNA markers or DNA barcodes (Larranaga et al. 2019, 2022), it provided a good framework for studying patterns of interrelationships and diversity within the Annonaceae, but it is of low resolution compared to chloroplast genome data (Senapati et al. 2023). The chloroplast genome database of Annonaceae has been upgraded and enhanced through the use of high-throughput genome sequencing technology and sequence analysis methods, A total of 14 chloroplast genomes of Annonaceae have been published on the National Center for Biotechnology Information (NCBI, <https://www.ncbi.nlm.nih.gov>), it accounted for 0.58% (Annonaceae more than 2,400 species), and the molecular evolution of chloroplasts in Annonaceae based on codon level has not been reported. In this study, the chloroplast genomes of 14 species were analyzed to reveal the differences in sequence characteristics and influencing factors of CUB in chloroplast genomes of these 14 species. To provide a theoretical basis and reference for the development and utilization of Annonaceae resources, phylogenetic relationships among related species, gene sequence evolution, species adaptive evolution, and classification of Annonaceae, and provide support for further research and application.

## Materials and methods

### Comparative genomic analysis

Chloroplast genome data of 14 species were collected from the NCBI database, including: *Polyalthiopsis verrucipes*, *Fissistigma oldhamii*, *Fissistigma polyanthum*, *Uvaria macrophylla*, *Greenwayodendron suaveolens*, *Chieniodendron hainanense*, *Annona muricata*, *Annona reticulata*, *Cananga odorata*, *Monoon laui*, *Miliusa glochidioides*, *Artabotrys pilosus*, *Artabotrys hexapetalus*, *Polyalthia suberosa*. The basic genomic features such as the total length of the sequence, the IR region, and the lengths of the four basic regions were preliminarily investigated (Wong et al. 2023). The IR region's boundary expansion and contraction were analyzed and mapped using IRSCOPE (<http://irscope.shinyapps.io/irapp>) (Uckele et al. 2021). The large sequence repeats (LSR) Four possible types of sequence repeats (forward match, reverse repeats, complement repeats and palindromic repeats) were analyzed by the online software REPuter (<http://bibiserv.techfak.uni-bielefeld.de/reputer/>) with parameter settings: sequence identity greater than 90%,

hamming distance of 3, and length greater than 30 bp (Rawal et al. 2021). The dispersed repeat sequences are found using the Tandem Repeats Finder (TRF). The simple sequence repeat (SSR) which is composed of repeating units in tandem with a length of 1–6 bp, can be used as a molecular marker and has been widely used in the analysis of population genetics of species (Mehmood et al. 2020). The software MISA (<https://webblast.ipk-gatersleben.de/misa/>) was used to identify simple repeat sequences in this chloroplast genome. The broad values of the number of repetitions from mononucleotide to hexanucleotide were set in order as 10, 5, 4, 3, 3, 3. All sequences were eventually manually corrected, and redundant repetitive sequences were manually removed (Tyagi et al. 2020).

## Analysis of codon usage bias of chloroplast genes

### Genomic data collection and sequences selection

After obtaining the whole genome sequence of chloroplasts of 14 species, according to the annotation information of gene bank the coding sequences (CDSs) were extracted. To improve the accuracy of codon bias analysis, we screened the sequences with the BioEdit software (Abdullah et al. 2020). First, the short (the length of the CDSs is less than 300 bp) and repetitive gene sequences were eliminated from the extracted CDSs. Then, the CDSs length must be multiples of three. More importantly, each CDS had a corrected start codon (ATG) at the beginning and a stop codon (TAA, TAG, and TGA) at the end (Waswa et al. 2023).

### Calculation of codon related parameters and codon composition analysis

In this study, the CodonW1.4.2 software was used to analyze the CUB of 686 selected CDSs of 14 species. These include relative synonymous codon usage (RSCU), relative frequency of synonymous codon (RFSC), high-frequency codons (HF), effective number of codon (ENC), codon adaptation index (CAI), codon bias index (CBI), frequency of optimal codons (FOP), total amino acids (L\_aa), and GC content at the 3rd base (GC3s) of the synonymous codon.

RSCU, RFSC, HF, ENC, CAI, CBI and FOP can be used to evaluate genomic CUB (Sahoo and Rakshit 2022). RSCU was used to detect variations in the pattern of usage of all synonymous codons across, it was a statistical measure of the factual degree of the relative recurrence of each synonymous codon and can be viewed simply as the ratio of the actual frequency of codon usage to its expected frequency (Cao et al. 2023). If an RSCU value greater than 1 indicates that the use frequency of the corresponding codon is higher than the expected frequency; RSCU value greater than 2

indicates that the use frequency of the corresponding codon is extremely high; RSCU value less than 1 indicates that the use frequency of the corresponding codon is lower than the expected frequency. The RFSC refers to the ratio of the sum of a codon observed in a test to the whole sum of synonymous codons, which reflects the usage frequency of each synonymous codon. For HF, the following principles were executed: RFSC > 60%; or the codon's RFSC is greater than 0.5 times the average of synonymous codons (Chen et al. 2023). The sequences with ENC values less than 30 and more than 55 correspond to high and low expression genes, respectively. The lower ENC values indicate higher CUB (Morton 2021). The CAI value refers to the degree of conformity between the used frequency of synonymous codons and optimal codons in the coding region and takes a value between 0 and 1, the larger values indicate greater adaptation and higher levels of gene expression. When CBI = 1 indicates that only the optimal codon was used, while CBI less than 0 indicates that no optimal codons were involved. The FOP value is the ratio of the number of optimal codons to their synonymous codons, and the value ranges from 0 to 1, while the FOP value of 1 indicates that only the optimal codon is used and a FOP value of 0 indicates that no optimal codon is used (Zhang et al. 2018).

### Optimal codons analysis

The optimal codons were identified by the procedure reported in the literature with slight modifications (Chen et al. 2023). First, we screened the eligible chloroplast genes of the 14 species, sorted the individual genes of the 14 species according to the ENC value from high to low, and screened about 10% of the genes from the anterior and posterior ends to construct a low-expression gene pool (ENC value is large) and high-expression gene pool (ENC value is small). The difference between the RSCU values  $\Delta\text{RSCU}$  ( $\Delta\text{RSCU} = \text{RSCU}_{\text{high expression}} - \text{RSCU}_{\text{low expression}}$ ) was calculated by the software of CodonW 1.4.2 (Uckele et al. 2021). A codon that meets  $\Delta\text{RSCU}$  values was not less than 0.08 is a high-expression codon. Second, the RSCU value of the corresponding codon was greater than 1 is an HF codon. Finally, those that satisfy both of the above two conditions were recognized as the optimal codon.

### Comparative analysis of codon usage frequency

The frequency of codon usage is the thousandth ratio of the number of particular codons to the total number of codons of the encoded gene (Zhao et al. 2023). The codon usage frequency data of *Arabidopsis thaliana* (<http://www.kazusa.or.jp/codon/cgi-bin/showcodon.cgi?species=3702>), *Nicotiana tabacum* (<http://www.kazusa.or.jp/codon/cgi-bin/>

[showcodon.cgi?species=4097](http://www.kazusa.or.jp/codon/cgi-bin/showcodon.cgi?species=4097)), *Escherichia coli* (<http://www.kazusa.or.jp/codon/cgi-bin/showcodon.cgi?species=199310>), and *Saccharomyces cerevisiae* (<http://www.kazusa.or.jp/codon/cgi-bin/showcodon.cgi?species=4932>) were downloaded from the Codon Usage Database (<http://www.kazusa.or.jp/codon>). In case the proportion is less than or equal to 0.5, or it is greater than or equal to 2, it indicates a more prominent distinction in CUB between two living beings to the extent that it would be inappropriate to select for a gene heterologous expression receptor (Cao et al. 2023). In contrast, the proportion is between 0.5 and 2, which demonstrates that the CUB is exceedingly comparative, then it can be used as a receptor of heterogenic expression.

### ENC-plot analysis

For the analysis of the ENC-plot, we plotted scatter plots with GC3 as the horizontal coordinate and ENC as the vertical coordinate. The factor affecting codon preference were judged by the location of gene distribution (Wang et al. 2022). If the genes were distributed on or near the standard curve, the codon preference was mainly affected by mutation; if the genes were located farther below the standard curve, the codon preference was more affected by natural selection and other factors (Yengkhom et al. 2019).

### Parity rule 2 plot analysis

In this study, we interpreted the mutational pressure and natural selection on nucleotide composition in DNA double strands of 14 species by the parity rule 2 plot (PR2-plot) analysis with reference to the method of Wu (2023). We took  $G3/(G3+C3)$  as the horizontal coordinate and  $A3/(A3+T3)$  as the vertical coordinate. It was assessed whether base mutations impact nucleotide base variation based on the proportions of A, T, G, and C bases. The position where the A3 value is equal to the T3 value and item G3 value is equal to the C3 value is the centroid, which indicates that there is no bias between the 2 complementary DNA strands by mutational pressure and natural selection. If the values of G3 and C3 or A3 and T3 are close to each other, it means that the codon preference of the chloroplast genome is only affected by mutational pressure; if the values of G3 and C3 or A3 and T3 are more different, it means that the codon preference is mainly caused by natural selection.

### Neutrality plot analysis

Neutrality plot analysis is a method for quantitatively analyzing the effects of mutational pressure and natural selection on CUB (Tyagi et al. 2020). The mutations in synonymous

codons usually occur at base position 3, whereas mutations at base positions 1 and 2 are mutations in non-synonymous codons, and they have a lower mutation rate. In the neutral plot, the average of GC1 and GC2 of each gene (GC12) was used as the vertical coordinate, and the GC3s value as the horizontal coordinate, with each point representing 1 independent gene. If all points are diagonally distributed, it means that there is no significant difference in the bases at the 3 positions of the codon, and CUB was only affected by mutation pressure. If all points are not diagonally distributed, and the correlation between the GC12 and GC3s is very low, it means that natural selection is the main influencing factor of CUB (Guo et al. 2020).

### Correspondence analysis

The correspondence analysis (COA) was used to characterize potential effects on CUB. To illustrate the CUB of 14 species, the analysis was based on the 59 RSCU values (ATG, TGG, and 3 termination codons were excluded from the analysis) for each gene. We made scatter plots with axis1 and axis2 as horizontal and vertical coordinates, the codon usage pattern was determined based on the distribution of points, and the correlation analyses were performed for axis1, axis2, CAI, ENC, GC3s, and, L\_aa (Geng et al. 2022).

### Statistical analysis

The CodonW 1.4.2 software (<https://codonw.sourceforge.net/>) was used to analyze the codon characteristic parameters. The CUSP in-line program in EMBOSS (<https://www.bioinformatics.nl/cgi-bin/emboss/cusp>) was used to calculate each gene GC content and total GC content ( $GC_{all}$ ), codon position 1, 2, and 3 bases of GC content (GC1, GC2, and GC3), and T, C, A, and G content (T3, C3, A3, and G3) at the third base of the codon. SPSS 26.0 software and EXCEL 2021 software were used to analyze the correlation between the chloroplast genome codon using preference correlation indicators. The Figures in this manuscript were completed by the software of R programming language and origin. Microsoft Office Word was used to edit the entire manuscript.

## Results

### Comparative genomic analysis

#### The basic characteristic of chloroplast sequences and codon composition of 14 species

The chloroplast genome structures of the 14 species in this study were all typical tetrameric structures, containing two single-copy regions (LSC/SSC) and two inverted repeat regions (IRA/IRb). As shown in Table 1, among the results of the length of the chloroplast genomes of the 14 species, the *A. reticulata* was the longest (201,906 bp), the *G. suaveolens* was the shortest (159,031 bp), and the average length was 175,293 bp. The average length of the 14 species was 175,293 bp. The results of the LSC, SSC, and IR regions represented that the length of the LSC region ranged from 69,650 to 126,172 bp, the SSC region from 3014 to 37,606 bp, and the IR region from 14,018 to 129,242 bp. The range of the total genes was 260–392, the highest number of CDS sequences was found in *A. reticulata* (115) the least in *G. suaveolens* and *M. laui* (81). The difference in tRNAs quantity was only 4, and the number of rRNA was identical.

#### The IR/SC sector boundary analysis of 14 species

The results of the IR/SC sector boundary analysis in 14 species are shown in Fig. 1. There were significant differences in IR/SC boundaries among the 14 species, but there were similarities within the same genus or closely related species. We observed a large expansion of the IR region in *A. reticulata*, *A. muricata*, and *U. macrophylla*, resulting in a large number of genes entering the IR region, which were all more than 100,000 bp in length, and which may be the reason why their total genome length was longer than other

species. For the LSC/IRb boundary, *P. verrucipes*, *M. laui*, *M. glochidioides*, *P. suberosa*, and *A. hexapetalus* were all located in the gene between *rps19* and *rpl2*; *F. oldhamii* and *F. polyanthum* were located the gene between *trnI* and *ycf2*, all other species were different. The IRb/SSC boundary of *P. verrucipes*, *M. laui*, and *P. suberosa* were within *ycf1* and beyond 15 bp, and most species were close to the *ndhF* gene in the SSC region. For the SSC/IRA boundary, *P. verrucipes*, *M. laui*, *M. glochidioides*, and *P. suberosa* were within the *ycf1* gene, and *F. oldhamii* was within the *ndhF* gene; *A. reticulata* and *A. muricata* have located the gene between *rpl32* and *tRNA*. In addition, the IRA/LSC boundary of most species was adjacent to the *trnH* gene in the LSC region. The boundary of LSC/IRb and IRA/LSC had a lot in common with *F. oldhamii*, *F. polyanthum*, *G. suaveolens*.

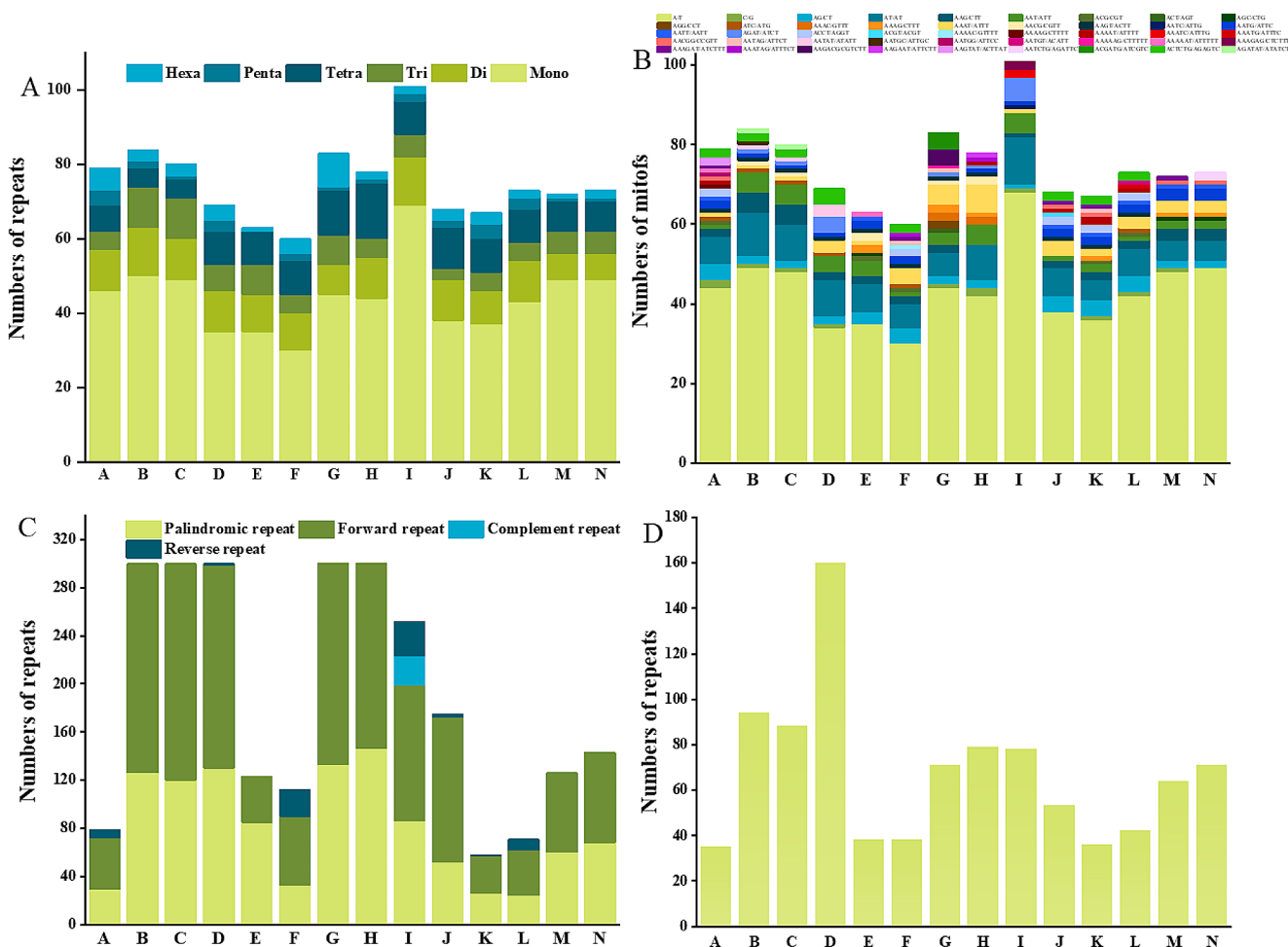
#### Repeat sequence analysis of the chloroplast genomes of 14 species

As shown in Fig. 2A, we detected 79, 84, 80, 69, 63, 60, 83, 78, 101, 68, 67, 73, 72 and 73 SSR loci in 14 species, respectively. Among them, the richest repeat type was mono-nucleotides (30–69), followed by di-nucleotides (7–13), tetra-nucleotides (5–15), tri-nucleotides (3–11), penta-nucleotides (0–4), and hexa-nucleotides (1–9), respectively. The A/T repeat motif accounted for a large proportion (30–68) among all of the repeat types, followed by AT/AT (5–12), but the other repeat motifs were very rare (Fig. 2B). The SSR motifs in different repeat class types of the 14 species were identified by REPuter (Fig. 2C), the number of repeats in each species ranged from 58 (*M. glochidioides*) to 300 (*F. oldhamii*, *F. polyanthum*, *U. macrophylla*, *A. muricata*, *A. reticulata*), with forward repeats (31–180) and palindromic repeats (24–146) being the most abundant types, followed by reverse repeats (0–29), the complement repeats are rare in these 14 species, only *C. odorata* has it.

**Table 1** The basic characteristic of chloroplast sequences of 14 species

Species	Genbank ID	Genome Size(bp)	LSC (bp)	SSC (bp)	IRs (bp)	Total Genes	CDS (before)	CDS (after)	tRNA	rRNA	
A	<i>P. verrucipes</i>	MW018366	159,965	89,030	18,987	51,948	267	84	47	37	8
B	<i>F. oldhamii</i>	MW829281	188,988	104,367	3043	81,578	307	106	51	38	8
C	<i>F. polyanthum</i>	MW829282	189,920	104,720	37,606	47,594	307	106	51	38	8
D	<i>U. macrophylla</i>	NC_041442	192,782	83,581	3741	105,460	311	105	47	38	8
E	<i>G. suaveolens</i>	NC_042164	159,031	126,172	18,841	14,018	299	81	46	37	8
F	<i>C. hainanense</i>	NC_043867	160,497	89,424	18,949	52,124	271	84	46	39	8
G	<i>A. muricata</i>	NC_052008	196,038	75,339	3105	117,594	384	111	50	38	8
H	<i>A. reticulata</i>	NC_052009	201,906	69,650	3014	129,242	392	115	49	40	8
I	<i>C. odorata</i>	NC_060837	167,946	83,620	20,310	64,016	271	84	50	39	8
J	<i>M. laui</i>	NC_061552	161,178	89,555	18,975	52,648	260	81	51	37	8
K	<i>M. glochidioides</i>	NC_062046	159,789	88,782	18,949	52,058	261	84	50	37	8
L	<i>P. suberosa</i>	NC_062666	159,408	88,566	19,000	51,842	264	85	50	37	8
M	<i>A. pilosus</i>	NC_063521	178,195	90,797	3098	84,300	287	96	50	38	8
N	<i>A. hexapetalus</i>	NC_063554	178,457	90,803	3066	84,588	316	90	48	36	8





**Fig. 2** Repeat sequence analysis of the chloroplast genomes of 14 species. A: The number of SSR in 14 species; B: The number of SSR motifs in different repeat class types; C: The number of four repeat

types; D: The number of dispersed repeat sequences; The number (A~N) of the 14 species are shown in Table 1

The results of the tandem repeats showed a total number of 947 tandem repeats for the 14 species, ranging from 35 (*P. verrucipes*) to 160 (*U. macrophylla*) (Fig. 2D).

the interim, the GC1 was upstairs GC2 and GC3, the GC3 was the lowest, there was a similar average of GC content among the 14 species (39.28-40.70%).

### Analysis of codon usage bias of chloroplast genomes

### Analysis of RSCU and RFSC

#### The basic characteristic of chloroplast sequences and codon composition of 14 species

In Table S1, the RSCU values of 686 CDSs of 14 species were analyzed and found that 30, 30, 30, 30, 30, 33, 30, 30, 29, 31, 28, 30, 30, 33, 32 codons of 14 species had RSCU values greater than 1, respectively. Among these codons with RSCU values greater than 1 in 14 species, the least number of codons ending in A/T was 83.3% and the greatest was 96.7%, while codons with RSCU < 1 are more biased to end in C/G, the results indicated that the codons preferred ending in A/T for the chloroplast genomes of 14 species. For codons with RSCU values greater than 1 there were 21 for 14 species, 12 ending in T, 8 ending in A, and 1 ending in G. By calculating the RFSC values of the 14 species, it was found that there were 17, 15, 14, 17, 18, 14, 14, 14, 15, 17,

As shown in Table 2, the 14 species chloroplast genome ENC was 48.600-51.041, CAI was 0.167-0.170, the mean value of CBI was -0.008, the mean value of FOC was 0.363, and the L-aa ranged from 15,177 to 22,220, the CAI value is low, while the ENC value is relatively high, indicating that the 14 species chloroplast genome CUB was weak. In 14 species the frequency of GC1, GC2, GC3, and the average of GC were below 50%, which indicated that the sequences of chloroplast genes preferred to end with A/T. In

**Table 2** The codon related parameters of 14 species

Species	ENC	CIA	CBI	FOC	L_aa	GC1/%	GC2/%	GC3/%	GCall/%
<i>P. verrucipes</i>	48.600	0.170	-0.080	0.360	17,649	47.49	39.31	31.18	39.33
<i>F. oldhamii</i>	50.345	0.168	-0.083	0.361	21,005	47.91	38.68	33.12	39.90
<i>F. polyanthum</i>	50.006	0.168	-0.083	0.362	22,220	47.25	38.21	32.76	39.41
<i>U. macrophylla</i>	50.082	0.168	-0.081	0.362	18,084	48.26	39.63	32.90	40.26
<i>G. suaveolens</i>	49.360	0.168	-0.078	0.365	15,177	47.93	40.29	31.25	39.82
<i>C. hainanense</i>	48.665	0.168	-0.081	0.361	16,966	47.71	39.83	32.04	39.86
<i>A. muricata</i>	51.036	0.169	-0.068	0.370	20,248	48.14	39.47	34.50	40.70
<i>A. reticulata</i>	51.041	0.167	-0.075	0.366	20,223	48.12	39.37	34.17	40.55
<i>C. odorata</i>	49.563	0.169	-0.081	0.362	20,750	47.33	38.60	32.41	39.44
<i>M. laui</i>	48.613	0.168	-0.081	0.362	18,793	47.60	39.29	31.29	39.39
<i>M. glochidioides</i>	48.749	0.169	-0.084	0.362	20,428	47.02	38.99	31.82	39.28
<i>P. suberosa</i>	48.956	0.169	-0.082	0.363	20,371	47.25	38.95	32.02	39.41
<i>A. pilosus</i>	49.599	0.168	-0.083	0.362	20,713	47.40	38.71	32.32	39.48
<i>A. hexapetalus</i>	49.819	0.170	-0.081	0.363	18,773	48.03	39.23	32.43	39.90

15, 15, 14, 14 HF in each of the 14 species, and 8 identical HF codons for 14 species (TGT, GAT, GAA, CAT, AAA, AAT, AGA, TAT).

### Determination of optimal codons of 14 species

As shown in Table 3 and Table S2, we screened a total of 244 optimal codons in 14 species with  $RSCU > 1$  and  $\Delta RSCU > 0.08$ , with 18,16,18,18,18,15,19,19,16,17,18,1

8,17,17 optimal codons in each species, respectively. All species shared 5 optimal codons (CGT, GTT, TGT, TTG, TTT). Among 244 optimal codons, 76 codons ended in A, 144 codons ended in T, 22 codons ended in G, and 2 codons ended in C, which indicated that the 3rd base of the optimal codon was biased towards A and T.

**Table 3** Optimal codon of chloroplast genome of 14 species

Species	Optimal codon number	Optimal codon
<i>P. verrucipes</i>	18	CTT, TAA, GCA, TAG, TGT, CCT, GTT, TCC, CGT, TTG, CGA, GTA, TTT, ACA, GAA, ATA, CAT, ACT
<i>F. oldhamii</i>	16	CTT, TAG, CGT, TTG, GCA, GTT, TGT, GGT, GCT, AAT, TTT, CGA, CCT, GTA, ACA, CAT
<i>F. polyanthum</i>	18	CTT, TCT, TGT, TAG, GTT, TTG, GCA, TTT, AGA, AAT, CGT, CCA, CGA, GCT, GGA, ACA, CAT, GTA
<i>U. macrophylla</i>	18	CGT, CCA, GTT, TGT, TTG, TAG, GGT, GGA, GCT, CTT, CCT, TTT, GTA, ACA, ATT, CAT, TAT, CGA
<i>G. suaveolens</i>	18	TAG, CTT, CGT, TGT, TCG, GTT, TTG, CCT, GCA, CAT, TAT, AGT, AAA, TTT, CGA, GGA, ATA, GTA
<i>C. hainanense</i>	15	CTT, TGT, AGT, TAA, CGT, CAT, CCT, GCA, TTG, GTT, GTA, ACT, TTT, TAT, ACA
<i>A. muricata</i>	19	CTT, CCA, CGT, ACA, GTT, TAA, GGT, GGA, TTT, CGA, GTA, TTG, GAA, GCT, AGA, TGT, ACT, ATT, AAT
<i>A. reticulata</i>	19	CTT, CGT, TGT, CCT, ACA, GTA, GAA, TTT, TTG, GGA, GGT, GTT, ACT, CCA, ATT, AGA, TCT, CGA, AAT
<i>C. odorata</i>	16	CGT, GGT, TGT, CTT, ACA, GCA, GTT, TTG, GTA, CAT, CCT, TAT, CCA, TCA, TTT, AGT
<i>M. laui</i>	17	CGT, GTT, GCT, TAG, CCT, TGT, ACA, TCG, TTT, TTG, ACT, CTT, CAT, GAA, TAT, GAT, GGA
<i>M. glochidioides</i>	18	CCT, CGT, CTT, GTA, TGT, GGT, AGT, CAT, TCT, TTT, ATT, TTG, GTT, GCT, ACA, CGA, GAA, ACT
<i>P. suberosa</i>	18	CTT, CGT, GTT, TGT, CAT, TTG, GGT, CCT, GCT, GTA, ACT, ACA, GAA, TTT, GGA, AGA, TAT, CGA
<i>A. pilosus</i>	17	TAA, CGT, TCT, CTT, TGT, TTG, GTT, CCT, GGT, GTA, GGA, TTT, CGA, CAT, ACT, CCA, TCC
<i>A. hexapetalus</i>	17	TTG, CGT, TAA, GTT, ACT, CCA, GGA, CCT, GCT, TGT, TTT, GTA, GGT, AGT, CGA, CAT, ATT



### Codon usage frequency analysis

The analysis of codon usage frequency of 14 species of Annonaceae and 4 four commonly used exogenous expression hosts illustrated that 14 species and *E. coli* had the largest differences, which were relatively small compared with the other 3 hosts (Table 4, Table S3). There were minor differences between *A. pilosus* and *S. cerevisiae*, the remaining 13 species exhibited the smallest differences from *N. tabacum*, even to the extent that *A. hexapetalus* and *N. tabacum* did not differ at all.

### ENC-plot analysis of 14 species

In the ENC-plot, it has been demonstrated that CUB on the standard curve is not affected by natural selection pressures, whereas codons distributed outside the standard curve are affected by natural or mutational selection pressures or other factors (Shahzadi et al. 2020). The results of this study demonstrated that 14 species partial genes (38%, 41%, 43%, 45%, 46%, 39%, 58%, 51%, 40%, 37%, 36%, 34%, 36%, and 42%) were located on or near the standard curve (Fig. 3; Table 5), which indicates that the actual ENC value of this partial locus is the same as the theoretical ENC value, and suggests that this partial gene's codon preference is not affected by the pressure of natural selection; whereas, among these, both *A. muricata* (58%) and *A. reticulata* (51%) were more than 50%, and the other part of the genes were farther away from the standard curve, which indicates that the actual ENC value of this part of the gene locus was farther away from the theoretical ENC value, then the codon preference of this part of the genes may be affected more by the pressure of natural selection.

In addition, in Fig. 3 the distribution of GC3s values for the 14 species is small (0.2244–0.401, 0.2308–0.4314,

0.2244–0.4142, 0.2308–0.3892, 0.2199–0.3774, 0.2308–0.4127, 0.2436–0.4513, 0.2308–0.4145, 0.2372–0.4032, 0.2179–0.3865, 0.2244–0.4047, 0.2244–0.4047, 0.2372–0.4207, 0.2372–0.4129), which also suggests that CUB was influenced by natural selection pressures.

### PR2-plot analysis of 14 species

The PR2-plot was used to reveal the influence of mutation and natural selection on CUB, by analyzing the utilization between A/T and G/C at the third codon position (Abdullah et al. 2020). As can be seen in Fig. 4, the distribution of bases A, G, C, and T in the four regions is uneven. In the vertical direction, most of the genes are located below the midline; in the horizontal direction, the number of genes on the right side of the midline is higher than that on the left side, but the difference in distribution is smaller than that in the vertical direction. Among them, the species *M. laui* has as many genes in the lower left and right region (quadrant III and IV), the species *P. verrucipes* and *C. hainanense* were the most abundant in the lower left region (quadrant III), and the remaining 11 species were the most abundant in the lower right region (quadrant IV). This suggests that base G frequency was greater than C, T was greater than A in the base composition of codon position 3 of 14 species, and further suggesting that most of the protein-coding genes in Annonaceae were subject to selective pressure.

### Neutrality plot analysis of 14 species

As can be seen in Fig. 5, the regression coefficients and correlation coefficients of *A. reticulata* and *A. muricata* were large (No. G, H), and their contributions of mutation pressure to codon preference reached 32.14% and 23.21%. These two species' codon preferences were most affected by mutational pressure and least affected by natural selection in 14 plants. On the contrary, *M. glochidioides* (No. K) was one of the 14 species whose codon preference was the least affected by mutational pressure and the most affected by natural selection. The correlation between GC12 and GC3 was low in all 14 species, ranging from 0.0007 to 0.1037, which suggests differences in the evolutionary patterns of GC12 and GC3. The correlation coefficient of GC12 and GC3 ranged from 0.0287 to 0.3214, which indicated that mutational pressure and natural selection affected the codon preference of the 14 species, but natural selection had a greater effect on codon preference. This result was in agreement with those of the ENC-plot and PR2-plot analyze, it was also consistent with edible legumes (Chakraborty et al. 2021).

**Table 4** The differences in codon usage frequency between 14 species and 4 hosts

	<i>A. thaliana</i>	<i>N. tabacum</i>	<i>E. coli</i>	<i>S. cerevisiae</i>
<i>P. verrucipes</i>	10	6	25	8
<i>F. oldhamii</i>	6	5	25	7
<i>F. polyanthum</i>	7	4	26	9
<i>U. macrophylla</i>	6	5	26	8
<i>G. suaveolens</i>	12	6	24	8
<i>C. hainanense</i>	11	4	24	8
<i>A. muricata</i>	7	5	25	9
<i>A. reticulata</i>	6	5	25	8
<i>C. odorata</i>	7	6	26	8
<i>M. laui</i>	11	5	25	9
<i>M. glochidioides</i>	9	5	27	10
<i>P. suberosa</i>	10	5	27	5
<i>A. pilosus</i>	9	5	19	3
<i>A. hexapetalus</i>	9	0	23	20

**Table 5** Frequency distribution of ENC ratio

Species		(-0.15, -0.05)	(-0.05, 0.05)	(0.05, 0.15)	(0.15, 0.25)	(0.25, 0.35)	Aggregate
P. verrucipes	Frequency	0	18	22	6	1	47
	Frequencies	0.00	0.38	0.47	0.13	0.02	1
F. oldhamii	Frequency	3	21	20	6	1	51
	Frequencies	0.06	0.41	0.39	0.12	0.02	1
F. polyanthum	Frequency	4	22	18	6	1	51
	Frequencies	0.08	0.43	0.35	0.12	0.02	1
U. macrophylla	Frequency	2	21	20	3	1	47
	Frequencies	0.04	0.45	0.43	0.06	0.02	1
G. suaveolens	Frequency	1	21	18	5	1	46
	Frequencies	0.02	0.46	0.39	0.11	0.02	1
C. hainanense	Frequency	1	18	21	5	1	46
	Frequencies	0.02	0.39	0.46	0.11	0.02	1
A. muricate	Frequency	1	29	14	4	2	50
	Frequencies	0.02	0.58	0.28	0.08	0.04	1
A. reticulata	Frequency	2	25	16	5	1	49
	Frequencies	0.04	0.51	0.33	0.10	0.02	1
C. odorata	Frequency	1	20	22	6	1	50
	Frequencies	0.02	0.40	0.44	0.12	0.02	1
M. laui	Frequency	0	19	24	6	2	51
	Frequencies	0.00	0.37	0.47	0.12	0.04	1
M. glochidioides	Frequency	2	18	21	8	1	50
	Frequencies	0.04	0.36	0.42	0.16	0.02	1
P. Suberosa	Frequency	2	17	23	7	1	50
	Frequencies	0.04	0.34	0.46	0.14	0.02	1
A. Pilosus	Frequency	4	18	20	6	2	50
	Frequencies	0.08	0.36	0.40	0.12	0.04	1
A. hexapetalus	Frequency	2	20	20	5	1	48
	Frequencies	0.04	0.42	0.42	0.10	0.02	1

### Correspondence analysis of 14 species

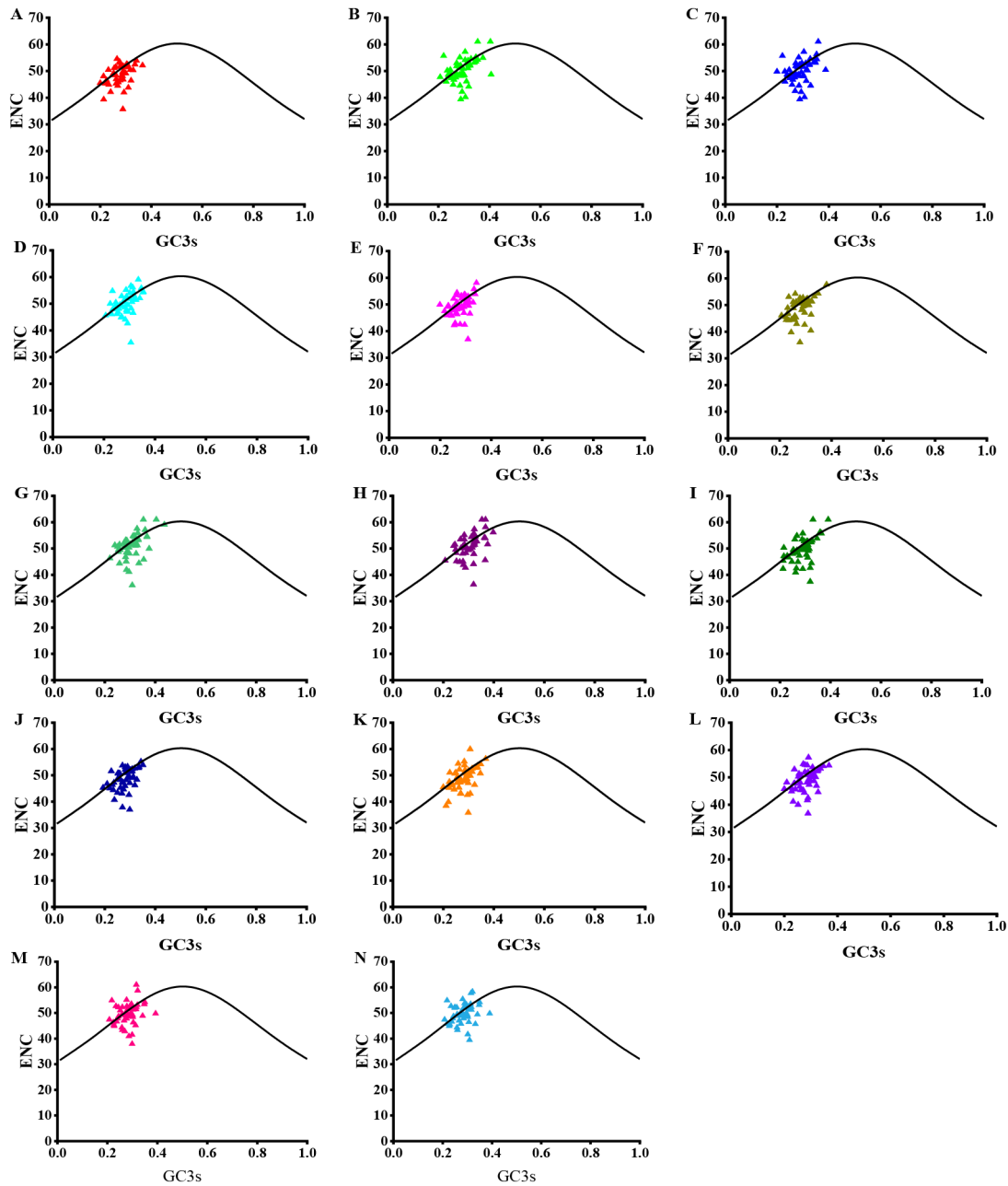
The COA of 14 species were shown in Fig. 6. The explainable variation in 1st axis in 14 plants accounted for 17.27%, 17.81%, 18.85%, 17.82%, 16.64%, 17.17%, 18.99%, 18.77%, 18%, 16.05%, 17.44%, 18.11%, 18.09%, and 17.26% of the total variation, respectively. The explainable variation in the 2nd, 3rd, and 4th axis of the 14 plants were smaller than the 1st axis. Thus, the 1st axis contributed the most to the codon bias, but the 1st axis represented only part of the variations in CUB (Zhang et al. 2018). Moreover, the different degrees of variation in codon usage bias of the four axes suggested that there is not a single factor influencing CUB, but that it may also be related to mutation, natural selection, the gene length and function, and other factors. In addition, the percentage of genes with different GC contents varied significantly among the 14 species,  $0.45 \leq GC \leq 0.60$  was 4.26%, 7.84%, 9.80%, 8.51%, 4.35%, 6.52%, 12.00%, 10.20%, 6.00%, 5.88%, 6.00%, 6.00%, 8.00%, 6.25%, respectively.

Furthermore, the correlation analysis was performed between axis 1, axis 2, and the codon index including CAI, ENC, GC3s, and L\_aa to explore the factors influencing

CUB (Fig. 7). We can see a correlation between axis 1 and CAI in all 14 species, and no correlated with ENC and GC3s, while axis 1 and L\_aa was positively correlated only for *G. suaveolens*, and not for any other species. For axis 2 and CAI, only the CAI of *F. oldhamii* was negatively correlated with axis 2; among ENCs, those of *P. verrucipes* and *M. glochidioides* were negatively correlated with axis 2, while those of *C. hainanense* and *P. suberosa* were positively correlated with axis 2; among GC3s, those of *M. glochidioides* was positively correlated with axis 2, the GC3s of *P. suberosa*, *A. pilosus*, and *A. hexapetalus* were positively correlated with axis 2, and other species were not correlated with axis 2; whereas, among the correlations between L\_aa and axis 2 of 14 species, only *F. oldhamii* and *G. suaveolens* had not correlate L\_aa and axis 2, and all other species had different degrees of correlation.

### Discussions

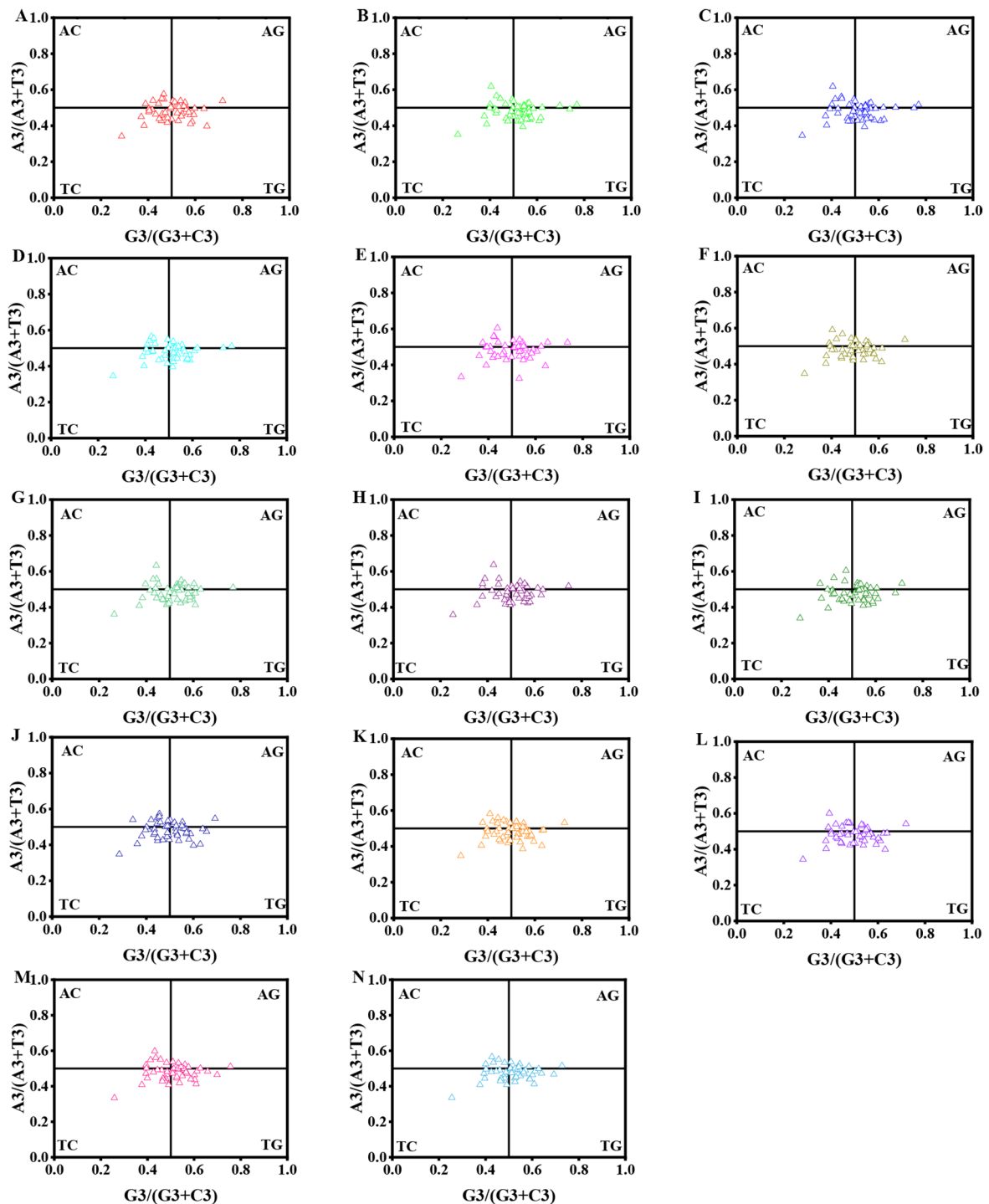
The chloroplast genome is the second largest genome and belongs to a relatively independent genome from the nuclear genome (Rehman et al. 2021). Although the chloroplast



**Fig. 3** ENC-plot analysis of chloroplast genomes of 14 species. The number (A~N) of the 14 species are shown in Table 1

genome is relatively stable, differences in structure, size, and number of genes may exist among families and species, this may be related to the contraction and expansion of the IR region and gene loss during genome evolution. Such differences can provide useful information for taxonomic, genetic, and ecological studies of plants (Konhar et al. 2019). In our study, the gene sizes of the 14 species varied greatly, ranging from 159,031 to 201,906 bp, which is consistent with the previously reported full-length range of chloroplast genome sequences, it demonstrates high conservation on Annonaceae chloroplast genes. The number of

chloroplast genes ranged from 260 to 392, which may be related to the metabolic pathways and ecological adaptations of different plants. The SSRs in chloroplast genomes were important molecular markers, that are widely used in the study of plant genetic diversity, species identification, and conservation biology. The SSRs of 14 species ranged between 63 and 101, and like most of the species mono-nucleotides had the highest number of repeats followed by di-nucleotides, tetra-nucleotides, tri-nucleotides, the penta-nucleotides, and hexa-nucleotides had the least number of repeats. Moreover, the A/T type is much larger than the G/C



**Fig. 4** PR2-plot of the chloroplast genomes of 14 species. The number (A~N) of the 14 species are shown in Table 1

type, which is consistent with the findings of other plants such as *Oryza* (Chakraborty et al. 2020), Theaceae (Wang et al. 2022) and Orchidaceae (Liu et al. 2023). In addition, the contraction, expansion, and deletion of the reverse repeat region can cause differences in the chloroplast genome, and the analysis of the IR/SC boundary and sequence variation

in Annonaceae found that there were significant differences in IR/SC boundaries among the 14 species, but there were similarities within the same genus or closely related species. In general, the whole genome is still more conservative, and the existence of different can provide a molecular

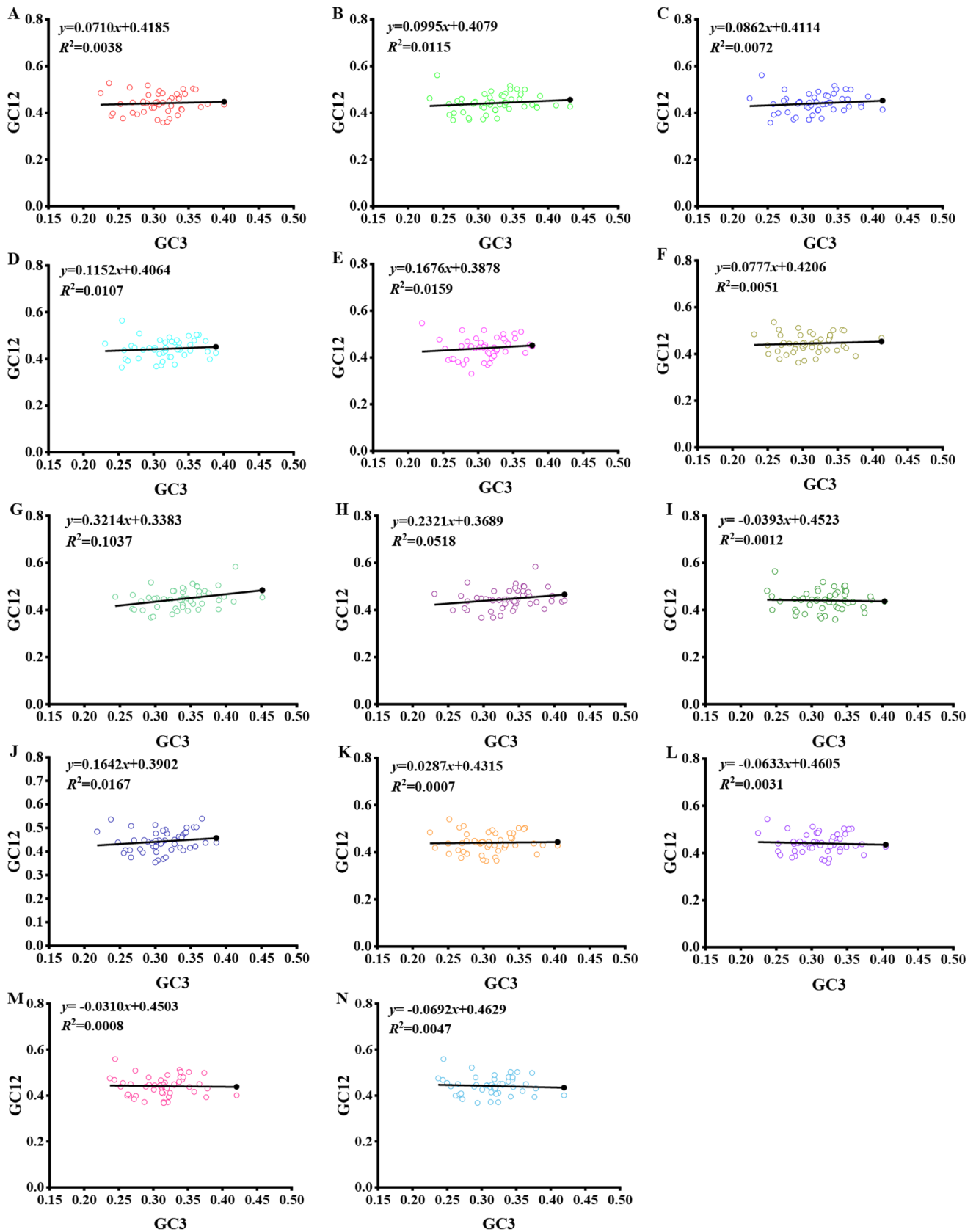
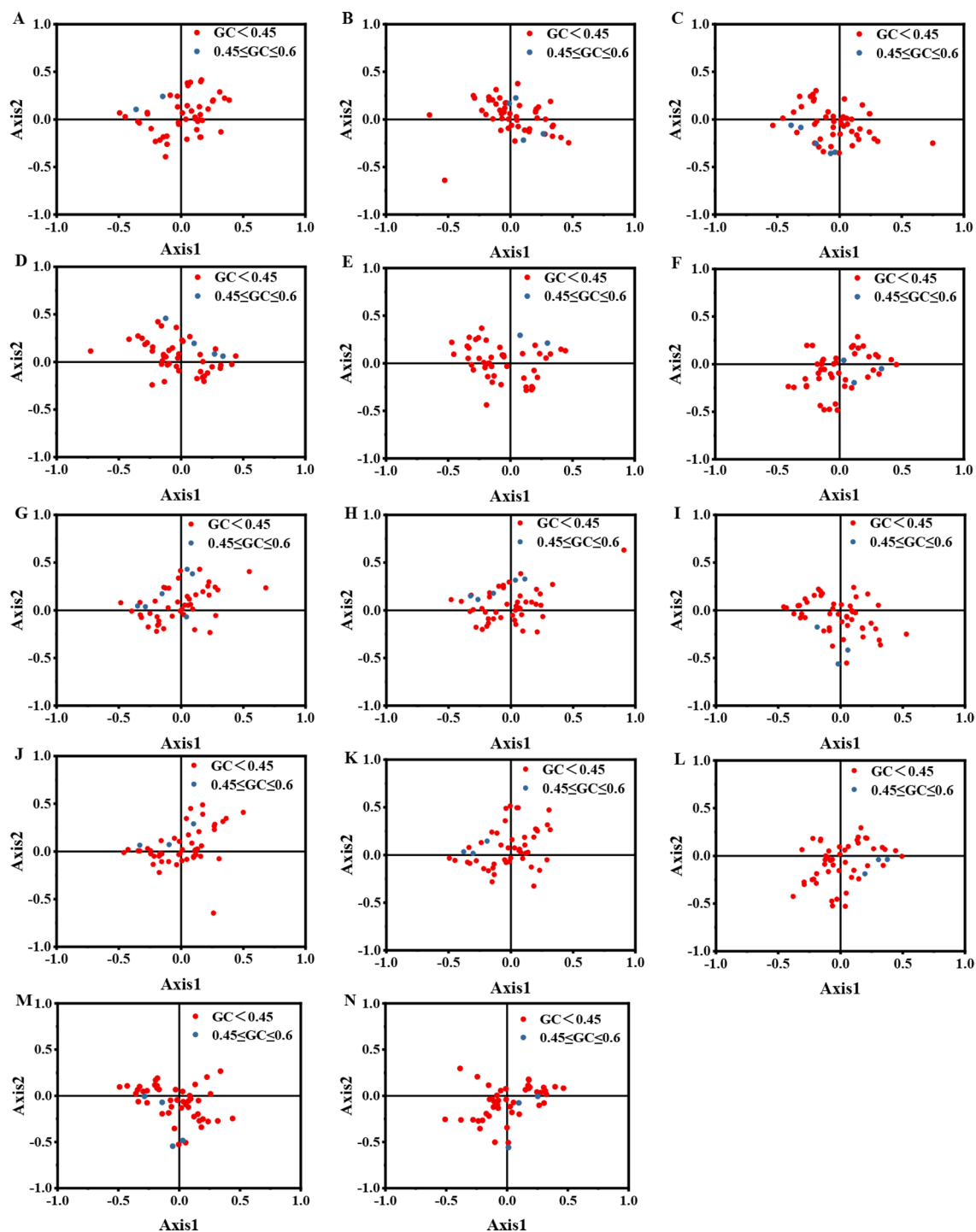


Fig. 5 Neutrality plot of chloroplast genomes of 14 species. The number (A~N) of the 14 species are shown in Table 1

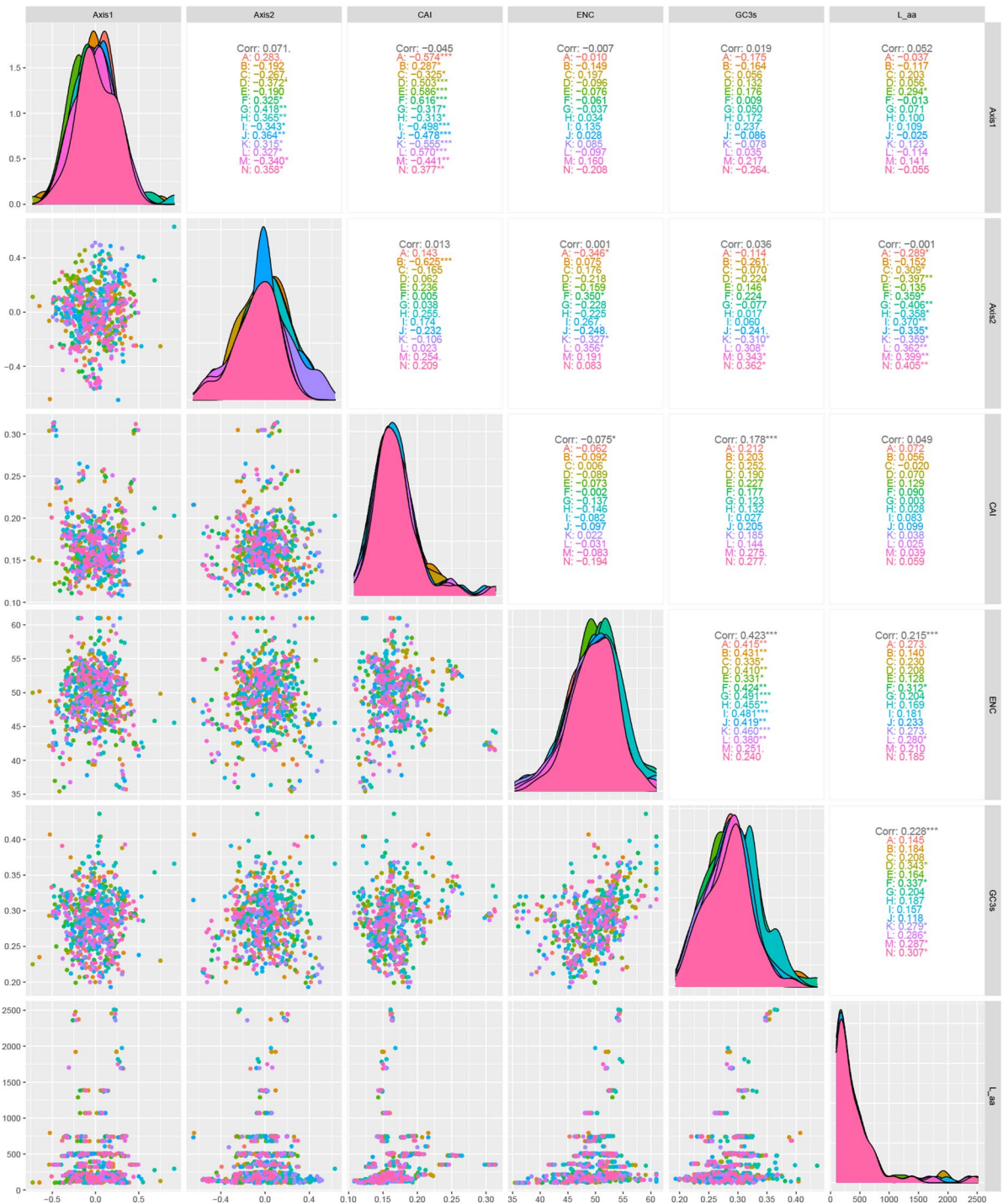


**Fig. 6** Correspondence analysis of chloroplast genomes of 14 species. The number (A~N) of the 14 species are shown in Table 1

basis for species identification and phylogenetic analysis in Annonaceae.

The ENC of 14 species were in the interval 48.600–51.041, with low gene expression levels, so CUB was relatively weak. In this study, the codon GC content of the chloroplast genomes of 14 species was lower than 50% and showed  $GC1 > GC2 > GC3$ , which also proved the

hypothesis “codons in higher plants tend to use A/T endings” (Campbell and Gowri 1990). PR2-plot analyses of 14 species represented that the CUB was not only affected by mutation but also by natural selection and other factors. Neutrality plot analysis and ENC-plot analysis also depicted that the CUB was largely influenced by natural selection, but little by mutational pressure during the evolution of



**Fig. 7** Correlation analysis of axis 1, axis2 and codon utilization index of chloroplast genomes of 14 species. The number (A~N) of the 14 species are shown in Table 1. \* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.001$

Annonaceae. Although both mutational pressure and natural selection can independently lead to the formation of CUB, and there are differences in the degree of impact and diversity among different species, the combination of these two mechanisms and the cumulative effect over a long period were primarily responsible for the formation of CUB. The

correlation analyses in 14 species demonstrated that the base composition of the 3rd position of the codon was different from the 1st and 2nd, it was related to the gene length and GC content. In addition, it was also found that the base composition of the 3rd position of the codon had a greater effect on the CUB, which suggested that there are differences in

the factors influencing CUB in different species. The study of optimal codons in organisms helps to maximize and improve translation speed and accuracy (Chi et al. 2023). A total of 244 optimal codons were screened in 14 species in this study, ranging from 15 to 19, it suggested that different species of Annonaceae have different strategies of optimal codon use in the process of adaptation to natural selection. A total of 220 codons ending in A/T, it was further confirmed that chloroplast genes of 14 species preferred to use NNA and NNT codons. Hershberg et al. (2008) revealed that the number of optimal codons was relatively small under purifying selection, and the number of optimal codons increased under high natural selection and mutation pressures. Therefore, it can be assumed that the chloroplast genomes of 14 species were not under purifying selection.

## Conclusions

In general, the genome characterization and codon usage patterns of chloroplast genomes were similar but not the same among Annonaceae. The chloroplast genome size of 14 species ranged from 159,031 to 201,906 bp, with LSC, SSC, IRa, and IRb tetrameric structure, and the size of each region varies greatly depending on the species. From the results, genome size, total number of genes, and CDS had some correlation. The SSRs ranged from 63 to 101, the richest in repeat type was mono-nucleotides, and the base composition preferred to use A/T; as with most species, the forward repeats and palindromic repeats were the most abundant types. The IR/SC sector boundary of chloroplast genomes of 14 species had obvious contraction and expansion. The frequency of GC was less than 50% in all 14 species. The codon-related parameters (ENC, CIA, CBI, FOC, and  $L_{aa}$ ) of 14 species varied significantly. These species had 21 identical codons with bias ( $RSCU > 1$ ), and there were 8 identical HF codons and 5 identical optimal codons. In ENC-plot and Neutrality plot analyses, CUB of all species except *A. muricata* and *A. reticulata* were relatively influenced by natural selection. The PR2-plot illustrated that base G frequency was greater than C, and T was greater than A. The correspondence analysis represented that codon usage patterns were different in Annonaceae.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s10142-024-01389-w>.

**Acknowledgements** We would like to thank the anonymous reviewers for their valuable comments on the manuscript.

**Author contributions** Author contribution X. H. and Y. L. were the designers and performers of the research, analyzed the data, and wrote the writing-original draft of the paper; F. M. was involved in the design

and analysis of the results of the experiment, provided the analytical tools, and guided the writing; Y. D. collected and organized of the data, reviewed the draft manuscript, and approved the final manuscript; M. S. curated and analyzed data, produced figures and charts, and guided the writing of the manuscript; S. Y. was involved in the analysis of results, wrote and reviewed the draft manuscript, and approved the final manuscript; H. L. was the conceptualizers and leaders of the project, and guided the study design, data analysis, writing, review and editing. All authors edited the manuscript and approve its submission in the current form.

**Funding** This research was funded by the project of Yuanmou Rural Revitalization Science and Technology Innovation Demonstration County, Grant number 202304BT090025, and the Revitalization of the Chuxiong industry technology leader training program.

**Data availability** No datasets were generated or analysed during the current study.

## Declarations

**Ethical approval** Not applicable.

**Competing interests** The authors declare that they have no conflict of interest associated with the work described in this manuscript.

## References

- Abdullah MF, Shahzadi I, Waseem S, Mirza B, Ahmed I, Waheed MT (2020) Chloroplast genome of *Hibiscus rosa-sinensis* (Malvaceae): comparative analyses and identification of mutational hotspots. *Genomics* 112(1):581–591. <https://doi.org/10.1016/j.ygeno.2019.04.010>
- Almutairi MM (2021) Analysis of chromosomes and nucleotides in rice to predict gene expression through codon usage pattern. *Saudi J Biol Sci* 28(8):4569–4574. <https://doi.org/10.1016/j.sjbs.2021.04.059>
- Campbell WH, Gowri G (1990) Codon usage in higher plant, green algae, and cyanobacteria. *Plant Physiol* 92(1):1–11. <https://doi.org/10.1104/pp.92.1.1>
- Cao Z, Yang L, Xin Y, Xu W, Li Q, Zhang H, Tu Y, Song Y, Xin P (2023) Comparative and phylogenetic analysis of complete chloroplast genomes from seven *Neocinnamomum* taxa (Lauraceae). *Front Plant Sci* 14:1205051. <https://doi.org/10.3389/fpls.2023.1205051>
- Chakraborty S, Yengkhom S, Uddin A (2020) Analysis of codon usage bias of chloroplast genes in *Oryza* species: Codon usage of chloroplast genes in *Oryza* species. *Planta* 252(4):67. <https://doi.org/10.1007/s00425-020-03470-7>
- Chakraborty S, Sophiarani Y, Uddin A (2021) Free energy of mRNA positively correlates with GC content in chloroplast transcriptomes of edible legumes. *Genomics* 113(4):2826–2838. <https://doi.org/10.1016/j.ygeno.2021.06.026>
- Chen J, Ma W, Hu X, Zhou K (2023) Synonymous Codon usage Bias in the Chloroplast genomes of 13 Oil-Tea *Camellia* Samples from South China. *Forests* 14(4):794. <https://doi.org/10.3390/F14040794>
- Chi X, Chen R, Zhang F, Chen S (2023) Comparative plastomes of species from Phrymaceae and Mazaceae: insights into adaptive evolution, Codon usage Bias and phylogenetic relationships. <https://doi.org/10.1139/gen-2023-0014>. *Genome*
- Dehlinger B, Jurs J, Lychuk K, Putonti C (2021) The dynamic Codon Biases: calculating prokaryotic codon usage biases. *Microb Genomics* 7(10):000663. <https://doi.org/10.1099/mgen.0.000663>



- Finkeldey R, Gailing O (2013) Chloroplasts. Brenner's Encyclopedia of Genetics (Second Edition). Academic Press. 525–527
- Geng X, Huang N, Zhu Y, Qin L, Hui L (2022) Codon usage bias analysis of the chloroplast genome of cassava. *South African Journal of Botany* 151(PA): 970–975. <https://doi.org/10.1016/J.SAJB.2022.11.022>
- Guo X, Wang Y, Wang S (2020) Complete chloroplast genome sequences from yellowhorn (*Xanthoceras sorbifolia*) and evolution analysis based on codon usage bias. *Int J Agric Biology* 24(4):676–684. <https://doi.org/10.17957/IJAB/15.1487>
- Hershberg R, Petrov DA (2008) Selection on codon bias. *Annu Rev Genet* 42:287–299. <https://doi.org/10.1146/annurev.genet.42.110807.091442>
- Jiang Y, Li BT (1979) Flora Reipublicae Popularis Sinicae. *Sci Press* 30(2):10–175
- Konhar R, Debnath M, Vishwakarma S, Bhattacharjee A, Sundar D, Tandon P, Biswal DK (2019) The complete chloroplast genome of *Dendrobium nobile*, an endangered medicinal orchid from north-east India and its comparison with related *Dendrobium* species. *PeerJ* 7:e7756. <https://doi.org/10.7717/peerj.7756>
- Larranaga N, Albertazzi FJ, Hormaza JI (2019) Phylogenetics of *Annona cherimola* (Annonaceae) and some of its closest relatives. *J Syst Evol* 57(3):211–221. <https://doi.org/10.1111/jse.12473>
- Larranaga N, Fontecha G, Albertazzi FJ, Palmieri M, Hormaza JI (2022) Amplification of Cherimoya (*Annona cherimola* Mill.) With chloroplast-specific markers: geographical implications on diversity and dispersion studies. *Horticulturae* 8(9):807. <https://doi.org/10.3390/HORTICULTURAE8090807>
- Lei JM, Liang ZR, Zhang H, Li YQ, Xue BE (2022) Advances in exploitation and utilization of wild fruit resources of Annonaceae in China. *J Fruit Sci* 39(01):121–130. <https://doi.org/10.13925/j.cnki.gsxb.20210301>
- Li BT, Michael GG (2012) Flora of China, vol 19. Science, pp 672–713
- Li N, Sun MH, Jiang ZS, Shu HR, Zhang SZ (2016) Genome-wide analysis of the synonymous codon usage patterns in apple. *J Integr Agric* 15(5):983–991. <https://doi.org/10.3390/ijms17081304>
- Liu L, Du J, Liu Z, Zuo W, Wang Z, Li J, Zeng Y (2023) Comparative and phylogenetic analyses of nine complete chloroplast genomes of Orchidaceae. *Sci Rep* 13(1):21403. <https://doi.org/10.1038/s41598-023-48043-2>
- Mehmood F, Abdullah, Shahzadi I, Ahmed I, Waheed MT, Mirza B (2020) Characterization of *Withania somnifera* chloroplast genome and its comparison with other selected species of Solanaceae. *Genomics* 112(2):1522–1530. <https://doi.org/10.1016/j.ygeno.2019.08.024>
- Morton BR (2021) Context-dependent Mutation dynamics, not selection, explains the Codon usage Bias of Most Angiosperm Chloroplast genes. *J Mol Evol* 90(1):17–29. <https://doi.org/10.1007/s00239-021-10038-w>
- Morton BR (2022) Context-dependent mutation dynamics, not selection, explains the codon usage bias of most angiosperm chloroplast genes. *J Mol Evol* 90(1):17–29. <https://doi.org/10.1007/s00239-021-10038-w>
- Nogueira BBAP, Mendes DS, Mendes RB, Pimenta S, Pereira MCT, Azevedo AM, Nietsche S (2022) Selection in half-sib progenies of *Annona squamosa* L.: an important step in the development of new cultivars. *Sci Hort* 302:111173. <https://doi.org/10.1016/j.scienta.2022.111173>
- Parmar R, Cattonaro F, Phillips C, Vassiliev S, Morgante M, Rajora OP (2022) Assembly and annotation of Red Spruce (*Picea rubens*) chloroplast genome, identification of simple sequence repeats, and phylogenetic analysis in *Picea*. *Int J Mol Sci* 23(23):15243. <https://doi.org/10.3390/IJMS232315243>
- Rawal HC, Borchetia S, Bera B, Soundararajan S, Ilango RVJ, Barooah AK, Sharma TR, Singh NK, Mondal TK (2021) Comparative analysis of chloroplast genomes indicated different origin for Indian tea (*Camellia Assamica* Cv TV1) as compared to Chinese tea. *Sci Rep* 11(1):110. <https://doi.org/10.1038/s41598-020-80431-w>
- Rehman U, Sultana N, Abdullah, Jamal A, Muzaffar M, Poczaï P (2021) Comparative chloroplast genomics in Phyllanthaceae species. *Diversity* 13(9):403. <https://doi.org/10.3390/D13090403>
- Sadhu L, Kumar K, Kumar S, Dass A, Pathak R, Bhardwaj A, Reddy VS (2023) Chloroplasts evolved an additional layer of translational regulation based on non-AUG start codons for proteins with different turnover rates. *Sci Rep* 13(1):896. <https://doi.org/10.1038/S41598-022-27347-9>
- Sahoo S, Rakshit R (2022) The pattern of coding sequences in the chloroplast genome of *Atropa belladonna* and a comparative analysis with other related genomes in the nightshade family. *Genomics Inf* 20(4):e43. <https://doi.org/10.5808/gi.22045>
- Senapati A, Chetri BK, Mitra S, Shelke RG, Rangan L (2023) Decoding the complete chloroplast genome of *Cissus quadrangularis*: insights into molecular structure, comparative genome analysis and mining of mutational hotspot regions. *Physiol Mol Biology Plants* 29(5):709–724. <https://doi.org/10.1007/S12298-023-01312-W>
- Shahzadi I, Abdullah, Mehmood F, Ali Z, Ahmed I, Mirza B (2020) Chloroplast genome sequences of *Artemisia maritima* and *Artemisia absinthium*: comparative analyses, mutational hotspots in genus *Artemisia* and phylogeny in family Asteraceae. *Genomics* 112(2):1454–1463. <https://doi.org/10.1016/j.ygeno.2019.08.016>
- Simmonds SE, Smith JF, Davidson C, Buerki S (2021) Phylogenetics and comparative plastome genomics of two of the largest genera of angiosperms, *Piper* and *Peperomia*. *Mol Phylogenet Evol* 163:107229. <https://doi.org/10.1016/J.YMPEV.2021.107229>
- Somaratne Y, Guan DL, Wang WQ, Zhao L, Xu SQ (2019) The complete chloroplast genomes of two *Lespedeza* species: insights into codon usage bias, RNA editing sites, and phylogenetic relationships in Desmodieae (Fabaceae: Papilionoideae). *Plants* 9(1):51. <https://doi.org/10.3390/plants9010051>
- Tyagi S, Jung JA, Kim JS, Won SY (2020) Comparative analysis of the complete chloroplast genome of mainland *Aster spathulifolius* and other *Aster* species. *Plants* 9(5): 568. <https://doi.org/10.3390/plants9050568>
- Uckele KA, Adams RP, Schwarzbach AE, Parchman TL (2021) Genome-wide RAD sequencing resolves the evolutionary history of serrate leaf *Juniperus* and reveals discordance with chloroplast phylogeny. *Mol Phylogenet Evol* 156:107022. <https://doi.org/10.1016/j.ympev.2020.107022>
- Wang Z, Cai Q, Wang Y, Li M, Wang C, Wang Z, Jiao C, Xu C, Wang H, Zhang Z (2022) Comparative Analysis of Codon Bias in the Chloroplast genomes of Theaceae Species. *Front Genet* 13:824610. <https://doi.org/10.3389/fgene.2022.824610>
- Wanichthanarak K, Nookaew I, Pasookhush P, Khoomrung S (2023) Revisiting chloroplast genomic landscape and annotation towards comparative chloroplast genomes of Rhamnaceae. *BMC Plant Biol* 23(1):59. <https://doi.org/10.1186/s12870-023-04074-5>
- Waswa EN, Mkala EM, Odago WO, Amenu SG, Mutinda ES, Muthui SW, Ding S, Hu G, Wang Q (2023) Comparative chloroplast genome analysis of *Sambucus* L. (Viburnaceae): inference for phylogenetic relationships among the closely related *Sambucus Adnata* Wall. Ex DC *Sambucus Javanica* Blume. *Front Plant Sci* 14:1179510. <https://doi.org/10.3389/fpls.2023.1179510>
- Wilson ME, Jensen GS, Haswell ES (2011) Two mechanosensitive channel homologs influence division ring placement in *Arabidopsis* chloroplasts. *Plant Cell* 23(8):2939–2949. <https://doi.org/10.1105/tpc.111.088112>
- Wong KH, Siu TY, Tsang SS, Kong BL, Wu HY, But GW, Hui JH, Shaw PC, Lau DT (2023) The Complete Chloroplast Genomes of Nine Smilacaceae Species from Hong Kong: Inferring Intra- and inter-familial phylogeny. *Int J Mol Sci* 24(8):7460. <https://doi.org/10.3390/ijms24087460>

- Wu P, Xiao W, Luo Y, Xiong Z, Chen X, He J, Sha A, Gui M, Li Q (2023) Comprehensive analysis of codon bias in 13 *Ganoderma* mitochondrial genomes. *Front Microbiol* 14:1170790. <https://doi.org/10.3389/fmicb.2023.1170790>
- Yengkhom S, Uddin A, Chakraborty S (2019) Deciphering codon usage patterns and evolutionary forces in chloroplast genes of *Camellia sinensis* var. *assamica* and *Camellia sinensis* var. *sinensis* in comparison to *Camellia pubicosta*. *J Integr Agric* 18(12):2771–2785. [https://doi.org/10.1016/S2095-3119\(19\)62716-4](https://doi.org/10.1016/S2095-3119(19)62716-4)
- Zhang R, Zhang L, Wang W, Zhang Z, Du H, Qu Z, Li X, Xiang H (2018) Differences in Codon usage Bias between Photosynthesis-related genes and genetic system-related genes of Chloroplast genomes in Cultivated and Wild *Solanum* species. *Int J Mol Sci* 19(10):3142. <https://doi.org/10.3390/ijms19103142>
- Zhang Y, Tian L, Lu C (2023) Chloroplast Gene expression: recent advances and perspectives. *Plant Commun* 4(5):100611. <https://doi.org/10.1016/J.XPLC.2023.100611>
- Zhao Y, Zhang X, Zhou T, Chen X, Ding B (2023) Complete chloroplast genome sequence of *Gynostemma guangxiense*: genome structure, codon usage bias, and phylogenetic relationships in *Gynostemma* (Cucurbitaceae). *Brazilian J Bot* 46(2):351–365. <https://doi.org/10.1007/S40415-023-00874-Z>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.